M2C: Energy Efficient Mobile Cloud System for Deep Learning

Kai Sun*, Zhikui Chen*, Jiankang Ren*, Song Yang†, Jing Li‡
*School of Software Technology, Dalian University of Technology, Dalian, China
e-mail: {elmsdk, zkchen, rjk}@{mail.dlut, dlut, mail.dlut}.edu.cn
†Delft University of Technology, Delft, Netherlands
e-mail: s.yang@tudelft.nl
‡Beijing branch, Tianjin Shenzhou Universal Data Technology Co. Ltd, Beijing, China
e-mail: lijing_0628@126.com

***Abstract*—with the number increasing of applications and services that are available on mobile devices, mobile cloud computing has drawn a substantial amount of attention by academia and industry in the past several years. When facing the most exciting machine learning applications such as deep learning, the computing requirement is intensive. For the purpose of improving energy efficiency of mobile device and enhancing the performance of applications through reducing execution time, M2C offloads computation of its machine learning application to the cloud side. We propose the prototype of M2C with the mobile side on Android, iPad and with the cloud side on the open source cloud: Spark, a part of the Berkeley Data Analytics Stack with NVIDA GPU. M2C's distinct set of varying computational tools and mobile nodes allows for thorough implementing distributed machine learning algorithm and innovative wireless protocols with energy efficiency, verifying the theoretical research and bringing the user extremely fast experience.**

I. Motivation

Energy is the non-replenishable resource in mobile devices. Currently, the energy requirement of a mobile device is supplied via lithium-ion battery that lasts only few hours if device is computationally engaged. Although the capacity of battery is increasing at the speed of 5 to 10% per year, the demand of computing and storage capability is increasing more rapidly. How to provide better user experience with constrained battery power supply is becoming more urgent than past years.

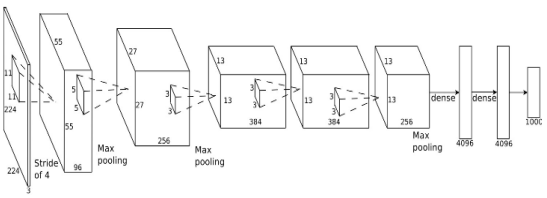

Fig. 1. Structure of Deep Learning [1]

The voice recognition applications of Baidu, Google and Bing such as Siri and Sogou voice assistant are mainly implemented based on deep learning. Deep learning is yielding new findings in the fields such as speech recognition and computer vision. It composed of multiple non-linear transformations [1]. Science writer John Markoff posits that deep learning would make surveillance technologies cheaper and more accessible, help marketers comb through data to identify consumer buying patterns, and may also pave the way for self-driving cars and robots that can replace human workers.

However, the deep learning algorithm is a relatively computing intensive algorithm. Therefore, the energy saving should be considered when implementing deep learning algorithm in mobile device. In addition, the constraining factors such as limitations in processing, storage and communication capabilities, energy supply, bandwidth, dynamic and unreliable network connectivity should also be considered for the deep learning applications comprising mobile devices. M2C is a mobile cloud system which could give the user extremely fast experience and save a quantity of energy of mobile device.

II. Structure of M2C

The deep learning system mainly comprises three parts: cloud side, data transmission channel and mobile device. Based on the constituent part of the deep learning system, the detailed structure of M2C is given as following.

- The overview of M2C

Spark[2] can run up to 100x faster than Hadoop MapReduce. Because, it uses a kind datasets which lets users persist data in memory. Cuda[4] could speed up every single node in cloud at least 5 to 6 times. Spark + cuda may give the mobile user the extremely fast experience.

The powerful computation capability of M2C will help mobile device execute most energy-intensive operation in it and enable near real-time response, this will save a quantity of energy for mobile device. A mobile node could upload the data to M2C, and then M2C gives back the processed result immediately to the node.

Besides these, we will optimize the most energy-intensive transmission mode: 3G transmission[6] for M2C: with an adaptive protocol to save transmission energy.

- The software stack of M2C

The structure of the cloud side of M2C is shown in Fig. 2. M2C is based on Apache Spark. Spark [2] which is an open source cluster computing system aims to make machine learning algorithm fast - both fast to run and fast to write. Spark

978-1-4799-3088-3/14/$31.00 ©2014 IEEE

was initially developed for two applications where placing data in memory helps: interactive data mining and iterative algorithms. This cloud could run on Amazon EC2 where user could run their cloud on the machines each one equipped with GPU.

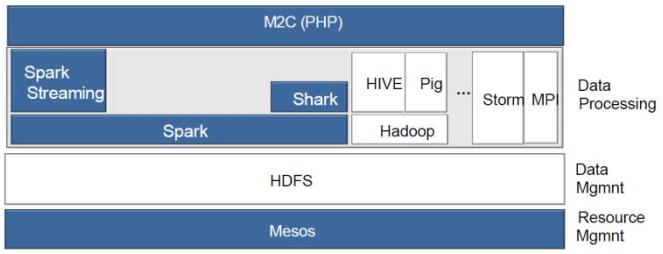

Fig. 2. Structure of M2C[2]

The foreground program of M2C is a popular php-based CMS (Content Management System) which called wordpress. The jQuery mobile technology can be easily integrated with PHP. More detailed description about the other part shown in Fig.2 can be found in [3]. The deep learning is a kind of iterative algorithms, thus the Spark-based cluster could improve the performance of this algorithm. For mobile device, the real-time responsive means reducing the delay time while saving energy.

Cuda toolkit could be implemented in the node of M2C cluster to leverage the performance of the whole system.

Some researchers have shown that they can comfortably train networks well with over 11 billion parameters — more than 6.5 times as large as the one reported in (Dean et al., 2012) (the largest previous network) , while using fewer than 2% as many machines. The Google brain uses GPU for their deep learning research, but Google didn't put the GPU cloud on line. We could implement this pattern on our own cluster or run the cloud in Amazon. The detail of GPU Cuda can be found in [4]. The data ready to be processed should be big enough, because Spark is designed for Big data just like Hadoop.

- The transmission channel structure of M2C

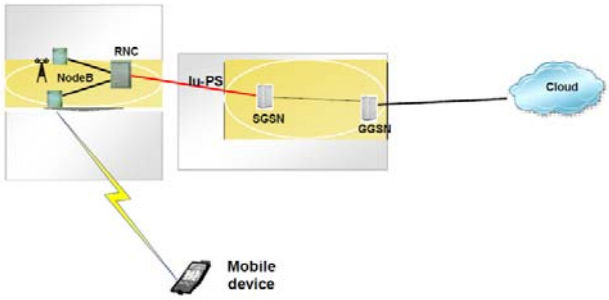

Fig. 3. Mobile device communicate with Cloud through WCDMA [5]

The transmission channel of M2C on 3G would be optimized. Here we give the transmission channel based on WCDMA for example. The transmission channel could schedule more uplink communication resource when the system running deep learning algorithm.

The parameter of transmission channel will be sent through a physical channel to RNC (Radio Network Controller) directly for QOS of mobile cloud computing instead of TCP/IP packet. In this way we may achieve significant reduction in the management overhead while maintaining the QOS (quality of service).

- The mobile node of M2C

The mobile node of M2C is implemented in android and iOS. And the android node uses Andriod NDK (native develop kit) technology to implement the adaptive selection protocol with C/C++ program language for cooperation of transmission channel. A program in the mobile node could record the energy consumption.

In mobile node, a self-defined application layer protocol is implemented and using it you could provoke a Cuda-based program in cloud side easily. In this pattern, the deep learning algorithm could be accelerated very rapidly.

HP LoadRunner 11.52 can be used to simulate a mobile user. Therefore, this program could simulate thousands of mobile nodes to visit the system. That we would test the performance of the cloud side and the protocols implemented in the system.

III. Conclusions

M2C considered almost every aspects of mobile cloud computing. M2C could let us research the cloud side, transmission channel, mobile node and every kind of theory at will. The energy consuming is not easy to simulate. And M2C could save the energy consumption in mobile node, and implement kinds of distributed deep learning algorithms running at very fast speed. It could also give user potential extremely fast experience.

ACKNOWLEDGMENT

I am grateful to Ivan Stojmenovic (fellow of IEEE and Canadian Academy of Engineering) and Baochun Li (senior member of IEEE) for their advices.

REFERENCES

[1] R. Salakhutdinov and G. Hinton, "An efficient learning procedure for deep Boltzmann machines," Neural Computation, vol. 24, pp. 1967-2006, 2012.

[2] C. Engle, A. Lupher, R. Xin, M. Zaharia, M. J. Franklin, S. Shenker, et al., "Shark: fast data analysis using coarse-grained distributed memory," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012, pp. 689-692.

[3] [Online].Available:http://spark.incubator.apache.org/talks/dev-meetup-dec-2012.pptx

[4] [Online].Available: http://docs.nvidia.com/

[5] [Online].Available: http://en.wikipedia.org/wiki/W-CDMA_(UMTS)

[6] X. Ma, Y. Cui, and I. Stojmenovic, "Energy efficiency on location based applications in mobile cloud computing: a survey," Procedia Computer Science, vol. 10, pp. 577-584, 2012